TAIL INDEX ESTIMATION WITH RANDOM BLOCK MAXIMA

STEFAN WAGER

Abstract. Using a backtesting framework, we develop a new estimator for the tail index of a distribution in the Fréchet domain of attraction. This estimator is equivalent to taking a U-statistic over a Hill estimator with two order statistics. The estimator presents multiple advantages over the Hill estimator. In particular, it has asymptotically $\mathcal{C}^\infty$ sample paths as a function of the threshold k, making it considerably more stable than the Hill estimator. The estimator also admits a simple and intuitive threshold selection heuristic that does not require fitting a second-order model.

Keywords: Heavy-tailed distributions, Hill estimator, infinite order U-statistics, tail index estimation, threshold selection.

1. Introduction

Researchers in multiple fields face a growing need to understand the tails of probability distributions, and extreme value theory presents tools which, under certain regularity assumptions, let us build simple yet powerful models for these tails. In the case of heavy tailed distributions, the setting of extreme value theory is as follows: Suppose our data is drawn from a distribution F, and assume that there is a constant $\gamma > 0$ and some slowly varying function L such that

$$1 - F(x) = L(x) \cdot x^{-\frac{1}{\gamma}}, \text{ with } \lim_{x\to\infty} \frac{L(ax)}{L(x)} = 1 \text{ for all } a > 0. \tag{1}$$

Then, F is in what is called the Fréchet domain of attraction.[1] If F satisfies this property (which most commonly used heavy-tailed distributions do), extreme value theory provides an elegant and concise description of the asymptotic properties of sample maxima of F. The only challenge is that this description relies on knowledge of the parameter γ, called the tail index of the distribution F. And, unfortunately, estimating γ from data is not always straightforward.

The literature on tail index estimation is quite extensive. One of the oldest and most widely used estimators is due to Hill [1975], who suggests estimating γ with a simple functional of the top $k + 1$ order statistics of the empirical distribution:

$$\hat{\gamma}_H := \frac{1}{k} \sum_{j=0}^{k-1} \log \left[\frac{X_{n-j,n}}{X_{n-k,n}} \right]. \tag{2}$$

Hill showed that $\hat{\gamma}_H$ converges in probability to $\gamma > 0$, provided the threshold sequence $k(n)$ is an intermediate sequence that grows to infinity slower than the sample size n. Hill's idea of using a functional of extreme and intermediate order statistics to estimate γ has received

Date: March 2, 2013.

[1] Extreme value theoretic methods are often discussed in the context of a larger family of distributions, characterized by a tail index $\gamma \in \mathbb{R}$. In this paper, however, we restrict ourselves to the Fréchet or heavy-tailed case with $\gamma > 0$.

considerable attention. Csörgő et al. [1985] suggest ways to adaptively weight the order statistics, while Dekkers et al. [1989] modify Hill's estimator so that it is also consistent for negative γ. More recently, Feuerverger and Hall [1999], Gomes et al. [2008], and others have worked on eliminating the asymptotic bias of the Hill estimator.

Nonetheless, tail index estimation remains quite challenging, especially for smaller samples on the order of a few hundred to a thousand points. Of course, many difficulties are inherent to the subject matter: Only a small fraction of any sample will be inside the tail of the underlying distribution, and so even large samples may contain very little information relevant to inference about this tail.

Other challenges, however, seem to arise from specifics of popular estimators. All estimators for γ require choosing a threshold at which the tail area of the distribution begins. Ideally, specifying a good threshold should be easy, and the estimate $\hat{\gamma}$ should not be sensitive to small changes in the threshold. Unfortunately, most commonly used estimators for γ do not reach this ideal. In the case of the Hill estimator—where the parameter k from (2) stands in for the threshold—the choice is far from innocuous:

- Inadequate choice of k can lead to large expected error. Small values of k have high variance, while large values of k usually have high bias. There is often an intermediate region for k where the estimator has fairly small expected error, but it is not always easy to find this region.
- The Hill estimator is extremely sensitive to small changes in k, even asymptotically: Mason and Turova [1994] show that the Hill estimator process converges in law to a modified Brownian motion. Thus, even within the 'good' region with low expected error, a minute change in k can impact the conclusions to be drawn from the model.

The problem of choosing the threshold k has been discussed, among others, by Beirlant et al. [2002], Danielsson et al. [2001], Drees and Kaufmann [1998], and Guillou and Hall [2001]. Most existing methods rely on fairly complicated auxiliary models: All but the last of the cited ones require either implicitly or explicitly fitting a difficult-to-fit second-order convergence parameter. As the method due to Guillou and Hall does not require fitting secondary parameters, we use it as our main benchmark in simulation studies. The problem of excessive oscillation of the Hill estimator has been discussed by Resnick and Stărică [1997], who recommend smoothing the Hill estimator by integrating it over a moving window. We are not aware of any guidance on how to automatically select k for this smoothed Hill estimator.

In this paper, we present a new estimator for γ which greatly simplifies the problem of threshold selection. Our estimator is based on a backtesting framework. It is well known that sample maxima from a distribution F satisfying (1) have the following property: If $X_1, ..., X_n$ are drawn independently from F, then as n goes to infinity, for some constants $a > 0$, $b \in \mathbb{R}$,

$$\frac{\max\{X_1, ..., X_n\}}{L(n) \cdot n^\gamma} \Rightarrow G_\gamma(ax + b),$$

where $G_\gamma(x)$ depends only on γ and $L(n)$ is an appropriately chosen slowly varying function. Noting this, we may suspect that when F has positive support,

$$\lim_{n\to\infty} n \cdot (\mathbb{E}[\log \max\{X_1, ..., X_n\}] - \mathbb{E}[\log \max\{X_1, ..., X_{n-1}\}]) = \gamma. \tag{3}$$

In Theorem 3.3, we show that this relation in fact holds under very mild conditions on F near 0.

Our estimator follows directly from this formula. We first estimate the quantities

$$\mathbb{E}[\log \max\{X_1, ..., X_s\}]$$

by subsampling our data without replacement, and then use (3) to obtain an estimate for $\hat{\gamma}$. Since this estimator operates by computing the average log maxima of random blocks, we call it the Random Block Maxima (RBM) estimator.

Our estimator behaves much like the Hill estimator; however, it addresses threshold selection much more naturally than the latter:

- The RBM estimator has asymptotically smooth sample paths as a function of k, and, even in modestly sized samples, does not suffer from small-scale instability in k.
- Thanks to its smoothness properties, the RBM estimator admits a simple and intuitive heuristic for threshold selection that does not require fitting a second-order model.

The estimator relates to backtesting in the sense that it tells us how much sample maxima would have increased with growing sample sizes in the past. For example, Petty et al. [2012] ask how much the value of the best deal seen by a venture capital firm might increase if the firm managed to expand the number of deals it evaluates by 10%. In this case, the quantity computed by the RBM estimator corresponds directly to the average increase in the log-value of the best sampled deal on permuted historical data.

The RBM estimator can be understood as belonging to two different frameworks of tail index estimation. The block maxima approach, which was often used in the early days of extreme value theory, aims to directly fit the distribution of fixed (e.g. yearly) blocks of data. In this light, the RBM estimator can be seen as a randomized method of moments estimator in the block maxima framework. Our estimator, however, can also be seen as an outgrowth of the more modern tail estimation paradigm started by the Hill estimator: As we will show, the RBM estimator can be constructed by taking a U-statistic over a Hill estimator with two order statistics. In other words, once we start subsampling the data, the block maxima and Hill estimation frameworks merge and lead to the RBM estimator.

In the next section, we outline how to use the RBM estimator, and apply it to a variety of datasets. After that, we study the theoretical properties of the estimator. We close with a simulation study which shows that, in terms of mean squared error (MSE), the RBM estimator is competitive with state-of-the-art threshold selection rules for the Hill estimator.

As the examples and the simulation study should make clear, the main advantage of the RBM estimator is not that it beats the state-of-the-art in tail index estimation by having low MSE. Rather, its strength lies in its stability and ease of use. Practitioners using the RBM estimator can get close to optimal estimates for γ by using an estimator $\hat{\gamma}(k)$ that is smooth in the tuning parameter k. We have already emphasized that this smoothness facilitates threshold selection, but the advantages do not stop there:

- The RBM estimator is stable enough in k that we can visually inspect the quality of the extreme value theoretic model and look for abnormal patterns that may indicate a failure of modeling assumptions by simply examining a plot of $\hat{\gamma}$ against k. In comparison, the corresponding curve for the Hill estimator is so noisy that it can be difficult to pick out any meaningful patterns with the naked eye.
- The smooth relationship between $\hat{\gamma}$ and k allows us to use labeled training data to choose k by supervised risk minimization – e.g. by running RBM on multiple datasets of the same size as our dataset of interest and with known γ, and then picking k with

the lowest prediction error. With the Hill estimator, the noise level is high enough that selection bias can easily overwhelm any true signal; however, with the RBM estimator the number of local minima to choose from is small and so the risk of problems related to selection bias is greatly reduced.

- With the RBM estimator, a small change in k will usually not produce a large change in $\hat{\gamma}$, and so it is more difficult for a marginally honest experimentalist to tune his choice of k in such a way as to get the value of $\hat{\gamma}$ he wants. Thus, in controversial situations, the RBM estimator may allow for less experimental bias than the Hill estimator.

Finally when paired with our threshold selection method, the RBM estimator allows us to get a point estimate for γ without having to fit a second-order model and without having to resort to manual threshold selection (for example, Coles [2001] recommends manually examining "a mean residual life plot" to select a threshold when estimating γ by maximum likelihood).

In other words, without compromising quality, our RBM estimator is easier to use and gives more stable estimates for γ than the Hill estimator, which is one of the most widely used tools for estimating the tail index of a heavy-tailed distribution.

2. Random Block Maxima

As described in (3), the RBM estimator for a given subsample size s is defined by

$$\hat{\gamma}_{RBM}(s) = s \cdot (M(s) - M(s-1)), \tag{4}$$

where $M(s)$ is the average log maximum of a subsample of size s drawn without replacement from the full sample of size n:

$$M(s) = \binom{n}{s}^{-1} \sum_{i_1 < \ldots < i_s} \max_{i_j} \{\log X_{i_j}\}. \tag{5}$$

Note that since we are interested in the behavior of sample maxima, we need to use resampling without replacement instead of with replacement. Otherwise, the presence of duplicate elements in our subsamples would bias our estimates $M(s)$ downwards.

To facilitate comparison between the Hill and RBM estimators, we do not parametrize our estimator directly in terms of the subsample size s, but use

$$k = \frac{2n}{s}, \tag{6}$$

which corresponds roughly to the degrees of freedom in the data used by the RBM estimator.

The RBM estimator has high variance for small k and potentially high bias for large k. More precisely, as shown in Theorem 3.3, the estimator has asymptotic variance

$$\lim_{n \to \infty} k(n) \operatorname{Var}[\hat{\gamma}_{RBM}(k(n))] = \gamma^2,$$

for any intermediate sequence $k(n)$, just like the Hill estimator. Asymptotic bias increases with k at a rate that depends on second-order parameters.

It is useful to plot $\hat{\gamma}_{RBM}(k)$ against k, which gives us an analog of a Hill plot. We have found such plots to be most informative when we plot k on a log scale rather than on a linear scale, as recommended by Drees et al. [2000]. Once we have computed $\hat{\gamma}_{RBM}(k)$ at multiple k, the problem becomes to choose which threshold $\hat{k}_{OPT}$ to use for estimating γ. A good

choice of threshold k should aim to simultaneously keep the bias and variance components small.

As we show in section 4, our estimator $\hat{\gamma}_{RBM}$ converges weakly to a $\mathcal{C}^\infty$ limiting process. In practice, $\hat{\gamma}_{RBM}$ is smooth enough as a function of k that we can reliably estimate its derivative in finite samples. This enables a particularly simple method for selecting a threshold $\hat{k}_{OPT}$ at which to report $\hat{\gamma}$.

We start by computing $\hat{\gamma}_{RBM}$ for subsample sizes $s = n, n-1, ..., 2$. By (6), these choices of s correspond to k-values $k_1 < k_2 < ... < k_{n-1}$. We then pick k using

$$\hat{k}_{OPT} = \operatorname{argmin}_m \left\{ \left(\frac{\hat{\gamma}_{RBM}(k_m) - \hat{\gamma}_{RBM}(k_{m-1})}{\log k_m - \log k_{m-1}} \right)^2 + \frac{\hat{\gamma}^2_{RBM}(k_m)}{2k_m} \right\}. \tag{7}$$

Roughly speaking, this heuristic aims to minimize the square of the derivative

$$\frac{\partial}{\partial \log k} \, \hat{\gamma}_{RBM}(k)$$

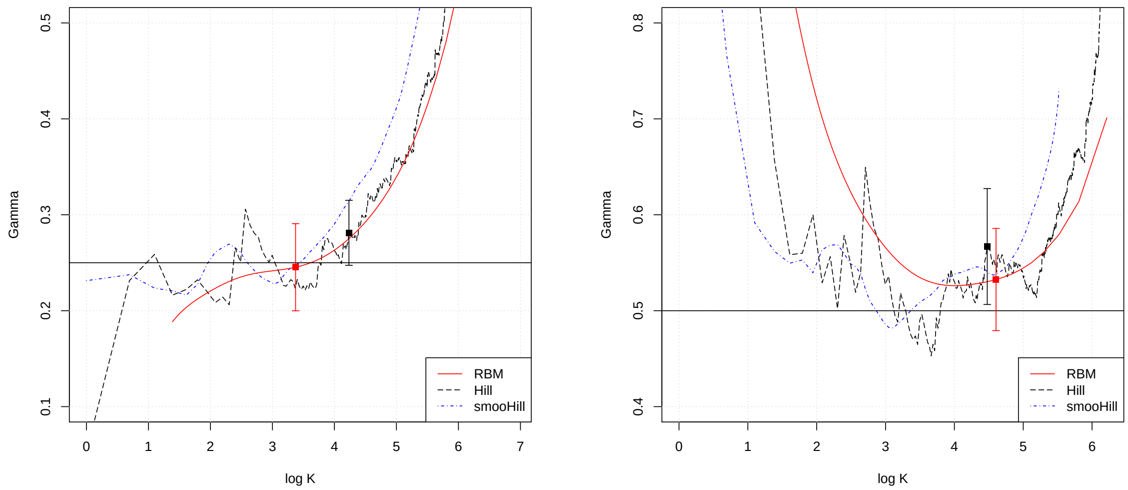
subject to a penalty term that decays as $1/k$. As argued in section 5, our choice of $\hat{k}_{OPT}$ aims to minimize possible bias in an empirical Bayes sense. We note that this threshold selection procedure is dependent on the smoothness properties of $\hat{\gamma}_{RBM}$. Attempting to use the same method with the Hill estimator $\hat{\gamma}_H$ would not lead to good results, since $\hat{\gamma}_H$ is not asymptotically differentiable as a function of k.

2.1. **Examples.** We showcase the RBM estimator by first applying it to two simulated datasets, and then using it to estimate the left hand tail index of daily stock market returns. The goal of these examples is to show how the RBM estimator can be used on real data; a more rigorous simulation study is given in section 6.

We compare the RBM estimator to both the Hill estimator and the smoothed Hill estimator (smooHill) proposed by Resnick and Stărică [1997]. There exist various heuristics for how wide a smoothing window to use for the smooHill estimator. We follow Resnick [2007] and average the Hill estimator on $(k, 2k]$ for each k. For the Hill estimator, we use the method from Guillou and Hall [2001] to automatically select k, while for RBM we use $\hat{k}_{OPT}$ from (7).

We begin by applying all three estimators first to 2000 datapoints drawn independently from a Student-t distribution with 4 degrees of freedom ($\gamma = 0.25$), and then to 500 datapoints from a Fréchet distribution with a shape parameter of 2 ($\gamma = 0.5$). In the case of the Student-t distribution, we discarded all negative datapoints (since all considered estimators involve taking logs of the datapoints), giving us an effective sample size of 992. Our results are given in Figure 2.1.

We observe that the RBM estimator oscillates much less than the Hill estimator or even the smooHill estimator (which has asymptotically $\mathcal{C}^1$ sample paths whereas the RBM estimator is asymptotically smooth). The instability of the Hill estimator is not benign: Around the selected threshold, a small change in k can shift the confidence interval for the estimator by a full standard deviation and potentially change conclusions drawn from the model. Thus, although the estimates given by the RBM estimator at the selected thresholds are not more accurate than those given by either the Hill or the smooHill at the same thresholds, they are much less ambiguous. This should be quite useful in applications, since the less ambiguous the answers given by an estimator are, the easier it is to evaluate convergence, and the less room there is for data dredging or other types of confusion.



(a) $N = 2000$ points drawn from a Student-t distribution with 4 df.

(b) $N = 500$ points drawn from a Fréchet distribution with $\chi = 2$.

FIGURE 1. Comparison of the RBM, Hill, and smooHill estimators on simulated data. The true value of γ is shown by a horizontal line, and the error bars are 1 standard deviation wide.

Next, we ran a similar comparison on 5 years of daily losses for the Dow Jones index. The dataset is described in Coles [2001], and is available online through the `R` package `ismev`. We only considered the 577 days on which the index lost value. Results are displayed in Figure 2.

Around the selected threshold, the RBM estimator is again substantially smoother than the Hill. We obtain estimates $\hat{k}_{OPT} = 33$ and a 95% confidence interval of $\gamma = 0.32 \pm 0.11$, while the Hill estimator paired with the threshold selection method from Guillou and Hall [2001] gives $\gamma = 0.35 \pm 0.09$. Both estimates are fairly close to the value $\hat{\gamma} = 0.29$ obtained by Coles [2001] using a maximum likelihood method. Coles, however, has a significantly wider confidence interval for this estimate (± 0.51), due in part to his use of weaker distributional assumptions that also allow for negative values of γ. Note that the confidence intervals for the Hill and RBM estimators may be somewhat optimistic here, as they assume independence of the data.

Finally we highlight a few cases where the RBM estimator as described here can fail, and show how to avoid these cases. First, the RBM estimator is somewhat computationally intensive. Our implementation can comfortably handle cases where n ranges in the low thousands; however, it becomes painfully slow when n approaches hundreds of thousands.[2] An easy way to avoid this problem without losing much information is to throw out all but the largest M datapoints (we usually take $M = 2'000$ or $10'000$). This speeds up the

[2] On our machine, we can run RBM with $n = 1'000$ in about 0.5 seconds and with $n = 10'000$ in roughly 10 seconds. When $n = 100'000$, the program runs in 5 minutes. We made reasonable attempts to optimize our code, but did not use any more sophisticated techniques like importing `C` libraries into `R`. The computational complexity of running the RBM estimator for all k is $\mathcal{O}(n^2 f(n))$, where $f(n)$ is the complexity of the procedure used to compute $n!$.

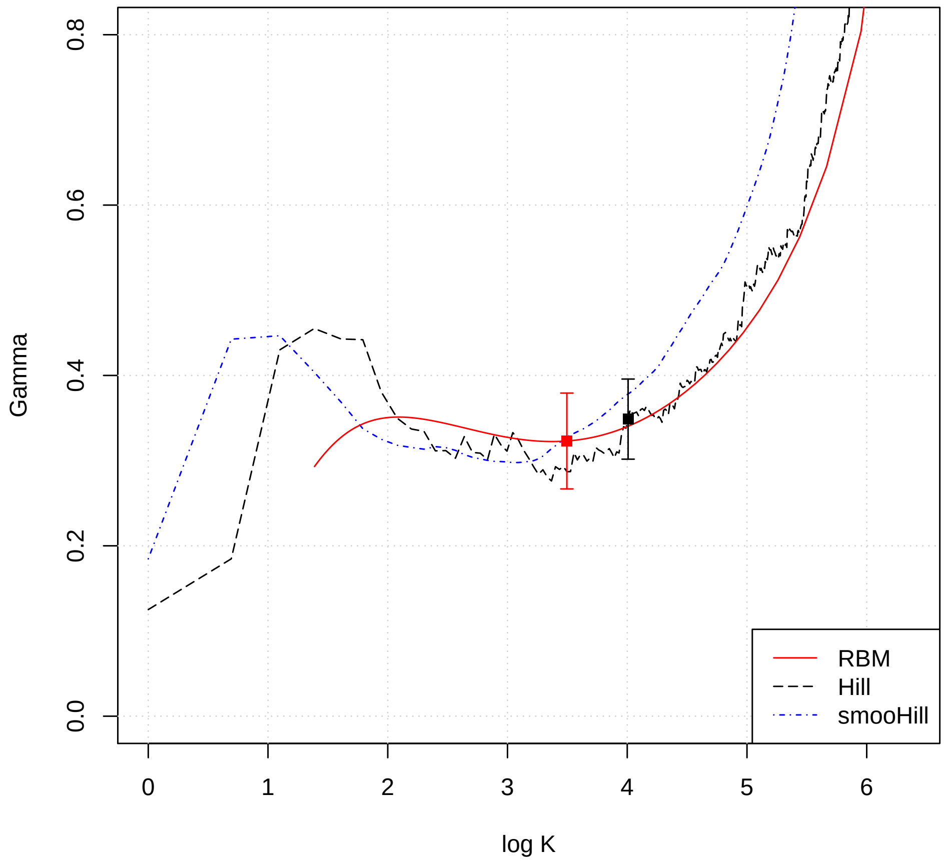


FIGURE 2. Comparison of RBM, Hill and smooHill estimators on Dow Jones daily negative returns. The error bars are 1 standard deviation wide.

algorithm a lot, and does not cost much in terms of accuracy because most of the information relevant to estimating γ is in the largest datapoints anyways.

Second, our threshold selection heuristic may fail if the data given to the RBM estimator is predominantly not from the tail of the distribution; this issue is discussed further in section 5. Again, a solution to this problem is to filter our data; in this case, we may want to throw out all the data that does not appear to be in the tail area we are trying to model.

3. Asymptotics of Random Block Maxima

We now move to theoretical results. The limiting distribution of the RBM estimator can largely be derived from the theory of U-statistics. A U-statistic is a multi-parameter generalization of a sample mean: Given data $X_1, ..., X_n$ and a symmetric s-parameter function f, the U-statistic over f is defined as

$$U_n\left(X_1, ..., X_n\right) := \binom{n}{s}^{-1} \sum_{\{I \subseteq \{1,...,n\}: |I| = s\}} f\left(X_I\right). \tag{8}$$

Such statistics have many desirable regularity properties. In particular, Hoeffding [1948] showed that when the underlying function f is held fixed, U-statistics are asymptotically normal with variance decaying as $1/n$.

As we have already stated earlier, our estimator $RBM_{k,n} := \hat{\gamma}_{RBM}(k)$ given in (4) can be described as a U-statistic over the Hill estimator. More precisely, for positive random variables $X_1, ..., X_s$ with s^{th} order statistics $X_{1,s} \leq ... \leq X_{s,s}$, let $H_1^{(s)}$ be the first Hill estimator on s datapoints

$$H_1^{(s)}(X_1, ..., X_s) := \log X_{s,s} - \log X_{s-1,s}. \tag{9}$$

We can then write $RBM_{k,n}$ as a U-statistic over $H_1^{(s)}$. The proof of the following lemma is given in the Appendix.

Lemma 3.1. *Let $X_1, ..., X_n$ be positive random variables with n^{th} order statistics $X_{1,n} \leq ... \leq X_{n,n}$. Then, the RBM estimator given in* (4) *is equivalent to*

$$RBM_{k,n} = \binom{n}{s}^{-1} \sum_{i_1 < ... < i_s} H_1^{(s)}(X_{i_1}, ..., X_{i_s}),$$

where k satisfies the relation $s = \lfloor 2n/k \rfloor$.

Expressing $RBM_{k,n}$ as a U-statistic enables us to leverage the extensive literature on the topic. Our problem, however, does not quite fall into the classical scope of U-statistics. Most of the literature assumes that the function f in (8) is fixed as n grows. But, in our case, the functions $H_1^{(s)}$ take a number of parameters that increases with n. Such a U-statistic is called an infinite order U-statistic. Although (as shown below) the classical asymptotic distributional results for U-statistics still hold in our case, the infinite order nature of the problem requires some additional work.

A common strategy for showing the asymptotic normality of a sequence of statistics $\{U_n\}$ is by approximating the U_n by their Hájek projections $\widehat{U}_n$. Suppose $X_1, ..., X_n$ are drawn from some known distribution, and let U_n be an n-parameter function. We then define its Hájek projection $\widehat{U}_n$ as

$$\widehat{U}_n := \mathbb{E}\left[U_n\right] + \sum_{i=1}^{n} \mathbb{E}\left[U_n - \mathbb{E}\left[U_n\right] | X_i\right]. \tag{10}$$

The advantage of studying such projections is that, when the X_i are *iid*, $\widehat{U}_n$ is a sum of independent random variables to which we can apply the central limit theorem.

When U_n is a U-statistic, U_n converges to $\widehat{U}_n$ in mean square under fairly general conditions.

Lemma 3.2. *Let $X_1, X_2, ...$ be independent and identically distributed random variables, and let $s(n)$ be a sequence such that $s(n) \leq n$ for all n. Moreover, let $g^{(n)}$ be a sequence of real-valued $s(n)$-parameter functions that are symmetric in their arguments, and let there be a constant C such that*

$$\text{Var}\left[g^{(n)}(X_1, ..., X_{s(n)})\right] \leq C$$

for all n. Then, taking U_n as a U-statistic over $g^{(n)}$

$$U_n = \binom{n}{s(n)}^{-1} \sum_{\{I \subseteq \{1,...,n\} : |I| = s(n)\}} g^{(n)}(X_I),$$

we find that

$$\mathbb{E}\left[\left(U_n - \widehat{U}_n\right)^2\right] = O\left[\left(\frac{s(n)}{n}\right)^2\right],$$

where $\widehat{U}_n$ *is defined as in* (10).

This lemma, whose proof is given in the Appendix, follows from the Efron-Stein ANOVA decomposition. We are now ready to prove our main result. As is common in extreme value theory, our result relies on a second-order convergence criterion. For an overview of this second-order condition, see e.g. de Haan and Ferreira [2006].

Theorem 3.3. *Let* $X_1, ..., X_n$ *be drawn* iid *from a distribution* F *satisfying the second-order condition*

$$\lim_{t\to\infty} \frac{\frac{U(tx)}{U(t)} - x^\gamma}{A(t)} = x^\gamma \frac{x^\rho - 1}{\rho}, \ \forall x > 0 \tag{11}$$

for some $\gamma > 0$*,* $\rho < 0$*, and a function* $A(t) \to 0$ *with constant sign. Here,* $U(t)$ *is the inverse quantile function* $U(t) = \inf\{x : \frac{1}{1-F(x)} \geq t\}$*. Moreover, suppose that* F *satisfies the technical condition*

$$\lim_{x\to 0} F(x) \cdot x^{-\frac{1}{\beta}} = 0 \text{ for some } \beta > 0, \tag{12}$$

and let $RBM_{k,n}$ *be the RBM estimator as described in Lemma 3.1.*

If $k(n)$ *is an intermediate sequence with* $k(n) \to \infty$ *and* $k(n)/n \to 0$ *such that*

$$\lim_{n\to\infty} \sqrt{k(n)} A\left(\frac{n}{k(n)}\right) = \lambda \text{ for some } \lambda \in \mathbb{R}, \tag{13}$$

then, for any $a > 0$*,* $RBM_{ak(n),n}$ *is asymptotically normal with*

$$\sqrt{k(n)}(RBM_{ak(n),n} - \gamma) \Rightarrow \mathcal{N}\left(\lambda\Gamma(1-\rho)\left(\frac{a}{2}\right)^{-\rho}, \frac{\gamma^2}{a}\right),$$

where Γ *is the gamma function. Moreover, for any* $a_1, ..., a_m > 0$*, the estimators* $RBM_{a_i k(n),n}$ *are aymptotically jointly normal with covariance*

$$\lim_{n\to\infty} k(n) \operatorname{Cov}[RBM_{a_i k(n),n}, RBM_{a_j k(n),n}] = \frac{2\gamma^2}{a_i + a_j}.$$

Proof. Let $s(n) = \lfloor 2n/k(n) \rfloor$ be the subsample block size. By Lemma 8.1,

$$\lim_{s\to\infty} \operatorname{Var}\left[H_1^{(s)}\right] = \gamma^2.$$

Thus, by Lemma 3.2, $RBM_{k(n),n}$ converges in mean square to its Hájek projection $\widehat{RBM}_{k(n),n}$, and

$$\lim_{n\to\infty} k(n) \cdot \mathbb{E}\left[\left(\widehat{RBM}_{k(n),n} - RBM_{k(n),n}\right)^2\right] = 0,$$

because $k(n) \cdot \left(\frac{s(n)}{n}\right)^2 \sim 4/k(n)$ converges to zero. Moreover, as in (26), for any $a > 0$ we can write this projection as

$$\widehat{RBM}_{ak(n),n} = \frac{s(n)}{an} \sum_{i=1}^{n} \mathbb{E}\left[H_1^{(s(n)/a)} | X_i\right] - \left(\frac{s(n)}{a} - 1\right) \mathbb{E}\left[H_1^{(s(n)/a)}\right]. \tag{14}$$

From Lemmas 8.2 and 8.3, we get that for any $a, b > 0$,

$$\lim_{n\to\infty} \frac{\mathbb{E}[H_1^{(s(n)/a)}] - \gamma}{A(2s(n))} = \left(\frac{a}{2}\right)^{-\rho} \Gamma(1-\rho), \text{ and}$$
$$\lim_{n\to\infty} s(n) \operatorname{Cov}\left[\mathbb{E}\left(H_1^{(s(n)/a)} | X_1\right), \mathbb{E}\left(H_1^{(s(n)/b)} | X_1\right)\right] = \frac{\gamma^2}{a^{-1} + b^{-1}},$$

the second of which implies, together with (14), that

$$\lim_{n\to\infty} k(n) \operatorname{Cov}\left[\widehat{RBM}_{ak(n),n}, \widehat{RBM}_{bk(n),n}\right] = \frac{2\gamma^2}{a+b}.$$

With these expressions in hand, we can conclude using the central limit theorem for triangular arrays and Slutsky's lemma that $RBM_{k(n),n}$ has the stated asymptotic distribution. □

The technical condition (12) is very weak, and can in practice be ignored. In an extreme value theoretic setup we usually care about very large values, whereas this condition only specifies the behavior of very small values. This condition trivially holds if F is supported on $[\varepsilon, \infty)$ for some $\varepsilon > 0$.

We end this section by noting that, by Lemma 8.1, even when F does not satisfy the the second-order condition for some $\rho < 0$, or when the sequence $k(n)$ does not satisfy (13), $H_1^{(2n/k(n))}$ still converges to γ in expectation. Thus, by a slight modification of the proof of Theorem 3.3, we find that, given any distribution F with tail index $\gamma > 0$, $RBM_{k(n),n}$ is consistent for γ along any intermediate sequence $k(n)$ provided F satisfies the technical condition (12).

4. The RBM Process

Our result from the previous section leads naturally to the definition of an RBM process. Under the conditions of Theorem 3.3 with some $\gamma > 0$ and $\rho < 0$, let $k(n)$ be an intermediate sequence such that, for some finite λ,

$$\lim_{n\to\infty} \sqrt{k(n)} A\left(\frac{n}{k(n)}\right) = \lambda.$$

Then, writing

$$X_n(t) = \sqrt{k(n)}\left(RBM_{tk(n),n} - \gamma\right), \tag{15}$$

our result in Theorem 3.3 implies that, for all $t_1, ..., t_m > 0$, the $X_n(t_i)$ are asymptotically jointly normal with

$$\lim_{n\to\infty} \mathbb{E}\left[X_n(t_i)\right] = \lambda\Gamma(1-\rho)\left(\frac{t_i}{2}\right)^{-\rho} \tag{16}$$
$$\lim_{n\to\infty} \operatorname{Cov}\left[X_n(t_i), X_n(t_j)\right] = \frac{2\gamma^2}{t_i + t_j}.$$

These mean and covariance equations can be used to define a Gaussian process, which we call the RBM process.

Definition 4.1. Given values $\gamma > 0$ and $\rho < 0$, let $X(t)$ be the Gaussian process on $\mathbb{R}_+^*$ satisfying the mean and covariance relations (16). The RBM process $R(\tau)$ is then defined by $R(\tau) = X\left(e^{\tau}\right)$ for $\tau \in \mathbb{R}$.

We define the RBM process on a log scale since this allows us to write down its properties more cleanly. This should not be too surprising, since the estimator as written in (4) is essentially a derivative $\frac{\partial M(s)}{\partial \log s}$ with s on a log scale. In a similar vein, Drees et al. [2000] show that the Hill process is most naturally plotted with k on a log scale.

The following lemma shows that the $X_n(t)$ do in fact converge in law to the process $X(t)$. The proof is given in the Appendix.

Lemma 4.2. *Let $X_n(t)$ be defined as in* (15) *under the conditions of Theorem 3.3, and let $X(t)$ be the auxiliary process from Definition 4.1 with the appropriate $\gamma > 0$ and $\rho < 0$. Then, the $X_n(t)$ converge weakly to $X(t)$ on compact intervals of $\mathbb{R}_+^*$ under the Skorokhod topology on $\mathcal{D}$.*

Our RBM process is analogous to the Hill process as discussed in Resnick and Stărică [1997]. These two processes, however, behave very differently. While the Hill process is equivalent to a modified Wiener process and so has continuous but non-differentiable sample paths, the RBM process has smooth sample paths.

Theorem 4.3. *There exists a modification of the RBM process defined in 4.1 that has $\mathcal{C}^\infty$ sample paths on $\mathbb{R}$ with probability one. Moreover, for any $\tau \in \mathbb{R}$, $R(\tau)$ and $R'(\tau)$ have joint distribution*

$$\begin{pmatrix} R(\tau) \\ R'(\tau) \end{pmatrix} \stackrel{d}{=} \mathcal{N}\left(\frac{\lambda\Gamma(1-\rho)e^{-\rho\tau}}{2^{-\rho}} \begin{pmatrix} 1 \\ -\rho \end{pmatrix}, \frac{\gamma^2}{e^\tau} \begin{pmatrix} 1 & -1/2 \\ -1/2 & 1/2 \end{pmatrix} \right).$$

Proof. It is well known [e.g. Loève, 1948] that, in order for a continuous-time stochastic process to have an almost surely $\mathcal{C}^\infty$ modification, it is sufficient for the covariance function $C(\tau_1, \tau_2)$ to be infinitely differentiable along the diagonal $s_1 = s_2$. We thus immediately get the desired smoothness result, since $\mathrm{Cov}[R(\tau_1), R(\tau_2)] = \frac{2\gamma^2}{e^{\tau_1} + e^{\tau_2}}$ is smooth on $\mathbb{R}^2$. The same result tells us that, for any $l, l' \in \mathbb{N}$,

$$\mathrm{Cov}\left[R^{(l)}(\tau), R^{(l')}(\tau)\right] = \frac{\partial^{l+l'}}{\partial u^l \partial v^{l'}} \mathrm{Cov}\left[R(u), R(v)\right] \Big|_{u=v=\tau},$$

which gives us the stated covariance result. The joint normality of R and R' and the expectation result follow directly from (16). □

In light of these results, we should expect the RBM estimator to have fairly smooth sample paths even for finite n. This is consistent with our observation in section 2.1 that the RBM estimator oscillates much less than either the Hill or the smooHill estimators.

5. Optimal Threshold Selection

Selecting an optimal tuning parameter k for the Hill estimator is a classic problem in extreme value theory. Both the Hill and the RBM estimators have high variance at small k, and may be quite biased at high k. A successful choice of k thus hinges on adequately balancing the bias and variance terms. Although the tuning parameter k is integrated fairly differently in the Hill and RBM estimators, a given choice of k has very similar effects on both estimators, and so our threshold selection heuristic should be read in light of the literature on optimal threshold selection for the Hill estimator.

Most approaches to selecting k require implicitly or explicitly estimating the second-order parameter ρ. Danielsson et al. [2001] and Hall [1990] suggest using various sub-sample

bootstraps to estimate the MSE-minimizing threshold in smaller samples. Transforming this small sample threshold into a full sample threshold, however, requires knowledge of ρ. Hall [1990] recommends just using $\rho = -1$, while Danielsson et al. [2001] use auxiliary bootstraps to estimate the correct transformation coefficient.

Drees and Kaufmann [1998] suggest a procedure based on a law of the iterated logarithm, which also requires fitting ρ. Finally, Beirlant et al. [2002] advocate plugging a consistent estimator for ρ into a formula for the optimal value of k given by Hall and Welsh [1985].

An alternative approach to threshold selection aims to stop just before the smallest value of k at which bias can be detected. Hill [1975] originally suggested picking k just before the log spacings between consecutive order statistics fail a test for exponentiality. This test, however, was shown by Hall and Welsh [1985] to be too lenient, and to produce estimates $\hat{\gamma}$ that were excessively biased. Guillou and Hall [2001] remedy this problem by developing a way to jointly test for bias among high-order log spacings. The approach advocated by Guillou and Hall [2001] does not require fitting ρ. This is a considerable benefit, since getting accurate estimates for ρ is not practical in many applications.

We suggest a threshold selection heuristic for the RBM estimator that is similar in spirit to this second class of alternatives, in that it aims to select a threshold just before significant bias starts to appear. However, instead of stopping just before bias can be detected at a given significance level, we aim to minimize possible bias in an empirical Bayes sense. Note that the following derivation of our threshold selection procedure is only intended as an informal motivation; the main argument for this procedure is that it is simple, intuitive, and appears to work in practice.

Consider the RBM process $R(t)$ discussed in section 4. From Theorem 4.3 we know that, if $\mathbb{E}[R(t)] = b(t)$ is the bias at t, then

$$R'(t) \overset{d}{=} \mathcal{N}\left(\rho b(t), \frac{\gamma^2}{2e^t}\right). \tag{17}$$

This suggests using $R'(t)$ as a test statistic for the hypothesis $b(t) = 0$.

A first approach to selecting the optimal threshold t would be to pick the first t at which we have to reject the null hypothesis $b(t) = 0$ at some significance level α. While this approach works decently, especially when $|\rho|$ is large, we found that an alternative empirical Bayes approach works even better.

Suppose that, for a fixed t, $b(t)$ is considered random with a uniform (improper) prior on $\mathbb{R}$. Then, using (17), we find that $b(t)$ has a posterior distribution

$$\mathcal{L}\left[b(t)|R'(t)\right] \overset{d}{=} \mathcal{N}\left(\frac{R'(t)}{\rho}, \frac{\gamma^2}{2\rho^2 e^t}\right),$$

and so

$$\mathbb{E}[b^2(t)|R'(t)] = \frac{2R'(t)^2 + \gamma^2 e^{-t}}{2\rho^2}.$$

We then select

$$\begin{aligned} \hat{t}_{OPT} &= \operatorname{argmin}_t \mathbb{E}[b^2(t)|R'(t)] \\ &= \operatorname{argmin}_t R'(t)^2 + \frac{\gamma^2}{2e^t}. \end{aligned} \tag{18}$$

We thus aim to select the value of t that gives us least cause to suspect bias, rather than the first t at which we must suspect bias. The heuristic given in (18) tends in practice to be

somewhat conservative, and selects thresholds t that have somewhat less bias and somewhat more variance than optimal from a MSE minimization point of view. Given, however, that the bias term is elusive whereas the variance term is easily estimated, such a tradeoff may not be so bad.

Our optimization heuristic in (18) takes the form of an intuitive penalized optimization problem. Broadly speaking, the procedure tries to select a point t such that R'_t is small, since low R'_t equates to low bias. However, low values of t are plagued by high variability, and so we penalize small values of t. This procedure seems to mimic the strategy a practitioner might use in selecting k from a Hill plot, and so we may hope that, even when the second-order condition does not hold or large third order effects are present, this heuristic will still give reasonable recommended thresholds.

We end this section on a note of caution: The relation (17) only holds in the tail region of the distribution. Thus, if we let k grow large enough that the RBM estimator starts to use substantial amounts of non-tail data, our heuristic can fail badly.[3] One way to avoid such a problem is, as discussed in section 2.1, to pre-filter our data and to only give the RBM estimator datapoints that are in the tail area of the distribution. For example, we might only use points that are above the mode of a coarse histogram of the data. Thankfully, such filtering should not cost us much, as the right-hand tail is the only part of the distribution that contains information that is relevant for estimating γ.

6. Simulation Study

In this section, we run simulations to test our RBM estimator against three other estimators for γ. The benchmark estimators are all threshold selection rules for the Hill estimator, and are described in detail in Beirlant et al. [2004]. We compare

- $\hat{\gamma}_{RBM}$: Our RBM estimator, with threshold selection implemented as in (7),
- $\hat{\gamma}_{BDGS}$: The plugin method from Beirlant et al. [2002],
- $\hat{\gamma}_{DK}$: The procedure based on a law of the iterated logarithm from Drees and Kaufmann [1998], and
- $\hat{\gamma}_{GH}$: The diagnostic for bias from Guillou and Hall [2001].

The distributions used for testing are given below. These distributions were also used for a simulation study in Beirlant et al. [2002].

- Fréchet(2) with distribution $F(x) = e^{-x^{-2}}$, $\gamma = 1/2$, $\rho = -1$. We drew $N = 200$ datapoints from this distribution.
- Burr(1, 0.5, 2) with distribution $1 - (1 + \sqrt{x})^{-2}$, $\gamma = 1$, $\rho = -1/2$. We drew $N = 500$ datapoints from this distribution.
- Student-t(6) with 6 degrees of freedom, $\gamma = 1/6$, $\rho = -1/3$. We drew $N = 500$ datapoints from this distribution.
- Log-Gamma(2, 1) with density $f(x) = x^{-2}\log(x)$, $\gamma = 1$, $\rho = 0$. We drew $N = 500$ datapoints from this distribution.

Simulation results are given in Table 1. All numbers were estimated using 4000 replications. Non-positive datapoints arising with the Student-t distribution were discarded, as discussed in section 2.1.

[3] To witness such a failure, one can try applying the RBM estimator on 10'000 datapoints drawn from a Student-t distribution with 2 degrees of freedom and a mean offset of +3. The heuristic from (18) will systematically pick a value of k that is much too large.

Table 1. Comparison of root mean squared error (RMSE) and bias for four estimators. Standard sampling errors ($\times 10^{-3}$) are indicated in parentheses.

Distribution		$\hat{\gamma}_{RBM}$	$\hat{\gamma}_{BDGS}$	$\hat{\gamma}_{DK}$	$\hat{\gamma}_{GH}$
Fréchet	RMSE	0.116 (2)	0.142 (4)	0.087 (1)	0.102 (1)
	Bias	0.011 (2)	-0.004 (2)	0.035 (1)	0.044 (1)
Burr	RMSE	0.334 (3)	0.442 (3)	0.344 (3)	0.382 (3)
	Bias	0.129 (5)	0.410 (3)	0.261 (4)	0.333 (3)
Student-t	RMSE	0.112 (1)	0.145 (1)	0.149 (1)	0.178 (1)
	Bias	0.074 (1)	0.130 (1)	0.134 (1)	0.168 (1)
Log-Gamma	RMSE	0.293 (2)	0.258 (3)	0.327 (2)	0.287 (2)
	Bias	0.215 (3)	0.182 (3)	0.301 (2)	0.238 (3)

We see that the RBM estimator is overall competitive with the other tested estimators in terms of MSE: RBM performs particularly well for both the Burr and the Student-t, and behaves reasonably for the rest. We note in particular that $\hat{\gamma}_{RBM}$ is substantially less biased than either $\hat{\gamma}_{GH}$ or $\hat{\gamma}_{DK}$ and somewhat less biased than $\hat{\gamma}_{BDGS}$ for the surveyed distributions. At equal MSE, having low bias may be advantageous since, as discussed earlier, variance terms are often easier to estimate than bias terms which depend on second-order parameters, and since systematic bias across multiple experiments may be hard to detect.

7. Conclusions

In this paper, we presented a new estimator for the tail index of a distribution in the Fréchet domain of attraction. The estimator arose from backtesting ideas, but can also be described as an infinite order U-statistic taken over the Hill estimator. The main advantage of our RBM estimator in comparison with existing methods lies in its stability and ease of use. While most commonly used estimators are extremely sensitive to small changes in the tuning parameter k, the RBM estimator is stable with respect to k. And, while most other estimators require either manually choosing the threshold or fitting a complicated auxiliary model for k, the RBM framework admits a simple, intuitive, and largely automatic heuristic for threshold selection. Although the results proved in this paper are asymptotic, we saw in section 2.1 that the advantages of the RBM estimator are apparent in finite samples.

More generally, this paper presents a new approach to constructing and finding the limiting distribution of tail index estimators. The asymptotic behavior of many classical estimators can be established using results from e.g. Drees [1998] on the convergence of tail empirical processes. In the present work, however, we take a different approach and study convergence using Hájek projections and infinite order U-statistics. There are multiple opportunities to tackle further problems in extreme value theory using similar methods. In particular, it should be possible to construct a bias-corrected version of the RBM estimator by mirroring ideas from Gomes et al. [2008], or to establish an RBM-type process which would permit estimation of a general tail index $\gamma \in \mathbb{R}$.

Acknowledgements

The author is grateful to Valérie Chavez-Demoulin, Bradley Efron, Julie Tibshirani, Suzanne de Treville, and especially Guenther Walther for discussions, comments, and helpful suggestions. This work was supported by a B. C. and E. J. Eaves Stanford Graduate Fellowship.

8. Appendix: Proofs

In the following results, we use the notation $U(t)$ for the inverse quantile function

$$U(t) = \inf\{x : \frac{1}{1 - F(x)} \geq t\}. \tag{19}$$

It can be shown [e.g. de Haan and Ferreira, 2006, Section 1.2] that the distribution F has extreme value index $\gamma > 0$ if and only if U is a regularly varying function of index γ, i.e. $\lim_{t \to \infty} U(tx)/U(t) = x^\gamma$ for all $x > 0$. We also use $[n]$ for the set $\{1, ..., n\}$.

Lemma 8.1. *Let $X_1, ..., X_m$ be drawn* iid *from a distribution F of strictly positive support with extreme value index $\gamma > 0$. Then the first Hill estimator $H_1^{(m)}$ from (9) converges in distribution to an exponential random variable with mean γ. Moreover, if there is a constant $\beta > 0$ such that*

$$\lim_{x \to 0} F(x) \cdot x^{-\frac{1}{\beta}} = 0,$$

then all moments of $H_1^{(s)}$ converge to the the corresponding moments of the limiting random variable. In particular,

$$\lim_{s \to \infty} \mathbb{E}\left[H_1^{(s)}\right] = \gamma \text{ and } \lim_{s \to \infty} \mathrm{Var}\left[H_1^{(s)}\right] = \gamma^2.$$

Proof. In terms of the inverse quantile function $U(t)$ from (19), we can write $X_k \overset{d}{=} U(Y_k)$, where the Y_k are drawn independently from a distribution with cdf $F_Y(y) = \frac{y-1}{y}$ for $y > 1$. We write $Y_{1,m} \leq ... \leq Y_{m,m}$ for the order statistics of the Y_k.

Since U is a regularly varying function of index γ, Potter's inequality [Potter, 1942] implies that, for any $\varepsilon > 0$, there is a t_0 such that, for all $t, \ tx \geq t_0$,

$$(1 - \varepsilon) x^{\gamma - \mathrm{sgn}[\log x] \cdot \varepsilon} < \frac{U(tx)}{U(t)} < (1 + \varepsilon) x^{\gamma + \mathrm{sgn}[\log x] \cdot \varepsilon}, \tag{20}$$

where sgn is the sign operator. Thus, since as in Lemma 8.4, $\lim_{m \to \infty} \mathbb{P}[Y_{m-1,m} < t_0] = 0$, we conclude that

$$\frac{U(Y_{m,m})}{U(Y_{m-1,m})} - \frac{Y_{m,m}}{Y_{m-1,m}} \to_p 0.$$

Now, we note that the $\log Y_k$ have standard exponential distribution *Exp*. By Rényi representation [Rényi, 1953], if $E_{1,m} \leq ... \leq E_{m,m}$ are order statistics of a standard exponential distribution, the $E_{k,m}$ are jointly distributed as

$$E_{k,m} \overset{d}{=} \sum_{l=1}^{k} \frac{E_l^*}{m - l + 1}, \text{ with } E_1^*, \ ..., \ E_m^* \sim Exp. \tag{21}$$

In particular, $E_{m,m} - E_{m-1,m}$ is exponentially distributed, and is independent from $E_{m-1,m}$. This implies our first claim:

$$H_1^{(s)} \stackrel{d}{=} \log\left[\frac{U(Y_{m,m})}{U(Y_{m-1,m})}\right] \Rightarrow \gamma \cdot \log\left[\frac{Y_{m,m}}{Y_{m-1,m}}\right] \stackrel{d}{=} Exp(\gamma).$$

To show convergence of the ν^{th} moment, we again use Potter's inequality, which implies that for any $\varepsilon > 0$ there is a t_0 such that

$$\begin{aligned} p_m(t_0) \cdot \mathbb{E}\left[(\log[1-\varepsilon] + (\gamma - \varepsilon) \cdot E)^\nu\right] &< \mathbb{E}\left[\log\left[\frac{U(Y_{m,m})}{U(Y_{m-1,m})}\right]^\nu ; Y_{m-1,m} > t_0\right] \\ &< p_m(t_0) \cdot \mathbb{E}\left[(\log[1+\varepsilon] + (\gamma + \varepsilon) \cdot E)^\nu\right], \end{aligned}$$

where $E \sim Exp$ and $p_m(t_0) = \mathbb{P}[Y_{m-1,m} > t_0]$. We recall that $p_m(t_0) \to 1$, and so in order to obtain convergence of moments it suffices to show that

$$\lim_{m\to\infty} \mathbb{E}\left[\log\left[\frac{U(Y_{m,m})}{U(Y_{m-1,m})}\right] ; Y_{m-1,m} < t_0\right] = 0;$$

this follows from the second part of Lemma 8.4, since the technical condition near 0 holds by hypothesis. □

Lemma 8.2. *Let $X_1, ..., X_m$ be drawn from a distribution F satisfying the second-order condition*

$$\lim_{t\to\infty} \frac{\frac{U(tx)}{U(t)} - x^\gamma}{A(t)} = x^\gamma \frac{x^\rho - 1}{\rho}$$

for all $x > 0$, with some $\gamma > 0$, $\rho < 0$ and a positive or negative function $A(t)$ with $\lim_{t\to\infty} A(t) = 0$. Moreover, suppose there is a constant $\beta > 0$ such that

$$\lim_{x\to 0} F(x) \cdot x^{-\frac{1}{\beta}} = 0.$$

Then, writing $X_{1,m} \leq ... \leq X_{m,m}$ for the order statistics of X, we have, for any $\alpha > 0$, that

$$\lim_{m\to\infty} \frac{\mathbb{E}[\log X_{m,m} - \log X_{m-1,m}] - \gamma}{A(\alpha m)} = \frac{\Gamma(1-\rho)}{\alpha^\rho}.$$

Proof. As in the proof of Lemma 8.1, we write $X_k \stackrel{d}{=} U(Y_k)$ where the Y_k have cdf $F_Y(y) = \frac{y-1}{y}$ for $y \geq 1$. Since $A(t) \to 0$, the stated second-order condition is equivalent to

$$\lim_{t\to\infty} \frac{\log U(tx) - \log U(t) - \gamma \log(x)}{A(t)} = \frac{x^\rho - 1}{\rho}$$

for all $x > 0$. By Drees [1998], there exists a function $A_0(t) \sim A(t)$ (and so without loss of generality $A_0(t) = A(t)$) such that for any $\varepsilon > 0$, there is a t_0 such that, for all $t > t_0$ and $x \geq 1$,

$$\left|\frac{\log U(tx) - \log U(t) - \gamma \log(x)}{A(t)} - \frac{x^\rho - 1}{\rho}\right| < \varepsilon x^{\rho+\varepsilon}. \tag{22}$$

For any $r < 1$ we find by Rényi representation (21) that

$$\mathbb{E}\left[\left(\frac{Y_{m,m}}{Y_{m-1,m}}\right)^r \big| Y_{m-1,m} \geq t_0\right] = \int_0^\infty e^{(r-1)x}\, dx < \infty,$$

and so, because $\rho < 0$, we find by plugging $t = Y_{m-1,m}$ and $tx = Y_{m,m}$ into (22) that for any $\delta > 0$ there is a t_0 such that

$$\lim_{m\to\infty} \mathbb{E}\left[\left(\frac{\log\left[\frac{U(Y_{m,m})}{U(Y_{m-1,m})}\right] - \gamma\log\left(\frac{Y_{m,m}}{Y_{m-1,m}}\right)}{A(Y_{m-1,m})} - \frac{\left(\frac{Y_{m,m}}{Y_{m-1,m}}\right)^\rho - 1}{\rho}\right)^2 ; Y_{m-1,m} \geq t_0\right] < \delta. \tag{23}$$

We now move to the case $Y_{m-1,m} < t_0$. $A(t)$ must be regularly varying [e.g. de Haan and Ferreira, 2006, Section 2.3] with index ρ, and so by Karamata representation we can assume without loss of generality that $A(t)$ is continuous on $[0, \infty)$ and strictily positive or strictly negative; in particular, $A(t)$ is then bounded away from 0 for finite intervals. Thus, by Lemma 8.4, the expression on (23) now integrated over the set $Y_{m-1,m} < t_0$ converges to 0. From this we conclude that

$$\lim_{m\to\infty} \mathbb{E}\left[\left(\frac{\log\left[\frac{U(Y_{m,m})}{U(Y_{m-1,m})}\right] - \gamma\log\left(\frac{Y_{m,m}}{Y_{m-1,m}}\right)}{A(Y_{m-1,m})} - \frac{\left(\frac{Y_{m,m}}{Y_{m-1,m}}\right)^\rho - 1}{\rho}\right)^2\right] = 0.$$

Moreover, assuming without loss of generality that appropriate regularity conditions for $A(t)$ hold near $t = 0$, we can show along the lines of Lemma 8.1 that

$$\limsup_{m\to\infty} \mathbb{E}\left[\left(\frac{A(Y_{m-1,m})}{A(m)}\right)^2\right] < \infty, \text{ and } \lim_{m\to\infty} \mathbb{E}\left[\left(\frac{A(Y_{m-1,m})}{A(m)} - \left(\frac{Y_{m-1,m}}{m}\right)^\rho\right)^2\right] = 0.$$

Thus, using Cauchy-Schwarz, we establish that

$$\lim_{m\to\infty} \mathbb{E}\left[\frac{\log\left[\frac{U(Y_{m,m})}{U(Y_{m-1,m})}\right] - \gamma\log\left(\frac{Y_{m,m}}{Y_{m-1,m}}\right)}{A(m)} - \frac{\left(\frac{Y_{m,m}}{m}\right)^\rho - \left(\frac{Y_{m-1,m}}{m}\right)^\rho}{\rho}\right] = 0.$$

Finally, by Rényi representation we can write

$$\begin{aligned}\left(\frac{Y_{m,m}}{m}, \frac{Y_{m-1,m}}{m}\right) &\overset{d}{=} \left(\frac{m^{-1}}{1 - \exp[-\tilde{E}_{1,m}]}, \frac{m^{-1}}{1 - \exp[-\tilde{E}_{2,m}]}\right) \\ &\Rightarrow \left(\frac{1}{E_1}, \frac{1}{E_1 + E_2}\right),\end{aligned}$$

where E_1 and E_2 are independent standard exponential and the $\tilde{E}_{k,m}$ are exponential order statistics. Uniform integrability holds, and so

$$\lim_{m\to\infty} \frac{\mathbb{E}\log\left[\frac{U(Y_{m,m})}{U(Y_{m-1,m})}\right] - \gamma\mathbb{E}\log\left[\frac{E_1+E_2}{E_1}\right]}{A(m)} = \mathbb{E}\left[\frac{\left(\frac{1}{E_1}\right)^\rho - \left(\frac{1}{E_1+E_2}\right)^\rho}{\rho}\right],$$

from which the desired conclusion follows by calculus and the fact that $A(t)$ is regularly varying of index ρ. □

Lemma 8.3. *Let $X_{1,m} \leq ... \leq X_{m,m}$ be independent order statistics drawn from a distribution F with extreme value index $\gamma > 0$, satisfying*

$$\lim_{x\to 0} F(x) \cdot x^{-\frac{1}{\beta}} = 0, \textit{ for some } \beta > 0.$$

Then, writing

$$\Psi_m(X) = \mathbb{E}[\log X_{m,m} - \log X_{m-1,m} | X_1 = X],$$

we have:

$$\lim_{m\to\infty} m \operatorname{Var}[\Psi_m] = \frac{\gamma^2}{2}, \text{ and, more generally,}$$
$$\lim_{m\to\infty} m \operatorname{Cov}[\Psi_m, \Psi_{\alpha m}] = \frac{\gamma^2}{1+\alpha}$$

for all $\alpha > 0$.

Proof. For convenience, write $W_i = \log X_i$. For $W_1, ..., W_m$, and $\tilde{W}_1, ..., \tilde{W}_{m+1}$ independent of each other,

$$\begin{aligned}
\delta_m(\tilde{w}) :&= \mathbb{E}[\tilde{W}_{m+1,m+1} - \tilde{W}_{m,m+1} | \tilde{W}_1 = \tilde{w}] - \mathbb{E}[W_{m,m} - W_{m-1,m}] \\
&= \mathbb{E}[W_{m-1,m} - \tilde{w}; W_{m-1,m} < \tilde{w} < W_{m,m}] + \mathbb{E}[\tilde{w} - 2W_{m,m} + W_{m-1,m}; W_{m,m} < \tilde{w}] \\
&= \mathbb{E}[W_{m-1,m}; W_{m-1,m} < \tilde{w}] - 2\mathbb{E}[W_{m,m}; W_{m,m} < \tilde{w}] \\
&\quad + \tilde{w} \cdot (2\mathbb{P}[W_{m,m} < \tilde{w}] - \mathbb{P}[W_{m-1,m} < \tilde{w}]).
\end{aligned}$$

Our goal is to study the distribution of $\delta_m(\log X)$ when $X \sim F$. We now proceed by evaluating each of these terms separately. As in the proof of Lemma 8.2,

$$W_{m,m} - \log U(m) \Rightarrow -\gamma \log E_1 \text{ and}$$
$$W_{m-1,m} - \log U(m) \Rightarrow -\gamma \log(E_1 + E_2),$$

where the E_i are independent standard exponential random variables.

We can use the Potter bounds (20) and Lemma 8.4 to show that the sequences $W_{m,m} - \log U(m)$ and $W_{m-1,m} - \log U(m)$ are uniformly integrable. This enables us to find the moments of interest from the limiting distributions. First, for all $w \in \mathbb{R}$,

$$\begin{aligned}
&\lim_{m\to\infty} \mathbb{E}[W_{m,m} - \log U(m); W_{m,m} - \log U(m) < w] \\
&\quad = -\int_{e^{\frac{-w}{\gamma}}}^{\infty} \gamma \log(x) \cdot e^{-x} \, dx \\
&\quad = w e^{-e^{\frac{-w}{\gamma}}} - \gamma \Gamma\left(0, e^{\frac{-w}{\gamma}}\right),
\end{aligned}$$

where Γ is the partial gamma function. Similarly,

$$\begin{aligned}
&\lim_{m\to\infty} \mathbb{E}[W_{m-1,m} - \log U(m); W_{m-1,m} - \log U(m) < w] \\
&\quad = -\int_{e^{\frac{-w}{\gamma}}}^{\infty} \gamma \log(x) \cdot x e^{-x} \, dx \\
&\quad = w(1 + e^{\frac{-w}{\gamma}}) e^{-e^{\frac{-w}{\gamma}}} - \gamma \left[e^{-e^{\frac{-w}{\gamma}}} + \Gamma\left(0, e^{\frac{-w}{\gamma}}\right) \right].
\end{aligned}$$

Finally,

$$\begin{aligned}
&\lim_{m\to\infty} 2\mathbb{P}[W_{m,m} - \log U(m) < w] - \mathbb{P}[W_{m-1,m} - \log U(m) < w] \\
&\quad = 2\mathbb{P}\left[E_1 > e^{\frac{-w}{\gamma}}\right] - \mathbb{P}\left[E_1 + E_2 > e^{\frac{-w}{\gamma}}\right] \\
&\quad = \left(1 - e^{\frac{-w}{\gamma}}\right) e^{-e^{\frac{-w}{\gamma}}}.
\end{aligned}$$

Combining all our expressions, we find that

$$\lim_{m\to\infty} \delta_m(w + \log U(m)) = \gamma \cdot \left[\Gamma\left(0, e^{\frac{-w}{\gamma}}\right) - e^{-e^{\frac{-w}{\gamma}}}\right]. \tag{24}$$

It remains to find the distribution of

$$z_m := \exp\left[-\frac{\log X - \log U(m)}{\gamma}\right],$$

when X is drawn from F. Now,

$$\begin{aligned}\lim_{m\to\infty} m\mathbb{P}[z_m < \chi] &= \lim_{m\to\infty} m\mathbb{P}\left[X > U(m)\chi^{-\gamma}\right] \\ &= \lim_{m\to\infty} m\mathbb{P}\left[X > U\left(m\chi^{-1}\right)\right] \\ &= \chi,\end{aligned}$$

for any $\chi > 0$. Thus, if μ_{z_m} is the distribution of z_m, we find that $m \cdot \mu_{w_m}$ converges weakly to Lebesgue measure on compact intervals of $\mathbb{R}_+$.

Now, by construction, we see that the functions $g_m(w) = \delta_m(w + \log U(m))$ must be Lipshitz continuous with constant 1 (since changing $\tilde{W}_1$ by Δ can change $\tilde{W}_{m+1,m} - \tilde{W}_{m,m}$ by at most Δ), and so the g_m converge uniformly on compact intervals to g, where $g(w) = f\left(e^{\frac{-w}{\gamma}}\right)$ and

$$f(z) = \gamma \cdot \left[\Gamma(0, z) - e^{-z}\right]$$

is the limiting function from (24). We can then argue by weak convergence of the μ_{z_m} to Lebesgue measure and by uniform convergence of $|g_m - g|$ to 0 that:

$$\begin{aligned}\lim_{c\to\infty}\lim_{m\to\infty} m\mathbb{E}\left[\delta_m(\log X); \left|\log\left[\frac{X}{U(m)}\right]\right| \le c\right] &= \lim_{c\to\infty}\int_{e^{-c/\gamma}}^{e^{c/\gamma}} f(z)\, dz = 0, \text{ and} \\ \lim_{c\to\infty}\lim_{m\to\infty} m\mathbb{E}\left[\delta_m^2(\log X); \left|\log\left[\frac{X}{U(m)}\right]\right| \le c\right] &= \lim_{c\to\infty}\int_{e^{-c/\gamma}}^{e^{c/\gamma}} f^2(z)\, dz = \frac{\gamma^2}{2}.\end{aligned}$$

It now remains to show uniform integrability of $m\delta_m^2$. Consider the residuals

$$R_c = \lim_{m\to\infty} m\mathbb{E}\left[\delta_m^2(\log X); \left|\log\left[\frac{X}{U(m)}\right]\right| > c\right].$$

By dominated convergence, if any one of the R_c is finite, then $\lim_{c\to\infty} R_c = 0$. Thus, the R_c only have two possible limiting values: 0 or infinity. Now, by Hoeffding's inequality [Hoeffding, 1948], we know that

$$m\,\mathrm{Var}[\Psi_m] \le \mathrm{Var}[H_1^{(m)}]$$

for all m. Moreover, from Lemma 8.1, we know that $\mathrm{Var}[H_1^{(m)}] \to \gamma^2$. Thus,

$$m\mathbb{E}[\delta_m^2(\log X)] = m\,\mathrm{Var}[\Psi_m] \le \gamma^2$$

for all m. This implies that the R_c are also bounded by γ^2, and so must converge to zero; thus our stated result about variance holds.

More generally, for any $\alpha > 0$, we find that

$$\begin{aligned}
&\lim_{m\to\infty} m \operatorname{Cov}[\Psi_m, \Psi_{\alpha m}] \\
&\quad = \gamma^2 \int_0^\infty \left(\Gamma(0,x) - e^{-x}\right) \cdot \left(\Gamma(0,\alpha x) - e^{-\alpha x}\right)\, dx \\
&\quad = \gamma^2 \left[x\Gamma(0,x)\Gamma(0,\alpha x) - \frac{e^{-(\alpha+1)x}}{\alpha+1} \right]_{x=0}^\infty \\
&\quad = \frac{\gamma^2}{\alpha+1}.
\end{aligned}$$

□

Lemma 8.4. *Let $X_{1,m} \leq ... \leq X_{m,m}$ be independent order statistics drawn from a distribution F of strictly positive support with extreme value index $\gamma > 0$. Then, for any fixed k and finite C,*

$$\lim_{m\to\infty} \mathbb{P}\left[X_{m-k,m} < C\right] = 0.$$

Moreover, if there is a constant $\beta > 0$ such that

$$\lim_{x\to 0} F(x) \cdot x^{-\frac{1}{\beta}} = 0,$$

then, for any $\nu > 0$,

$$\lim_{m\to\infty} \mathbb{E}\left[\left| \log[X_{m-k,m}] \right|^\nu \cdot 1\{X_{m-k,m} < C\} \right] = 0. \tag{25}$$

Proof. As in the proof of Lemma 8.1, we write $X_k \overset{d}{=} U(Y_k)$. Because $U(t) \to \infty$, the first statement follows directly by applying the strong law of large numbers to $1\{Y_k > r\}$ for a properly chosen $r > 0$. To prove the second part, we see that

$$\mathbb{E}\left[\log[X_{m-k,m}]_+^\nu \middle| X_{m-k,m} < C\right] \leq \log[C]_+^\nu$$

is uniformly bounded, and we already know that $\mathbb{P}\left[X_{m-k,m} < C\right]$ converges to zero. The hard part of establishing (25) is thus to establish a uniform bound for $\mathbb{E}\left[\log[X_{m-k,m}]_-^\nu \middle| X_{m-k,m} < C\right]$.

Now, because $\lim_{x\to 0} F(x) = 0$,

$$\begin{aligned}
\lim_{x\to 0} F(x) \cdot x^{-\frac{1}{\beta}} = 0 &\iff \lim_{x\to 0} \left(\frac{F(x)}{1-F(x)}\right)^\beta \cdot x^{-1} = 0 \\
&\iff \lim_{y\to 0} y^\beta \cdot U(1+y)^{-1} = 0,
\end{aligned}$$

where we obtain the last equivalence by writing $y = \frac{F(x)}{1-F(x)}$.

Without loss of generality, we picked C such that $C = U(t_0)$ for some t_0. Because $U(t)$ is monotone increasing, we can find a constant L such that

$$(t-1)^\beta \cdot U(t)^{-1} \leq L$$

for all $1 < t \leq t_0$. (Without loss of generality, let $L = 1$.) This implies that

$$\begin{aligned}
\mathbb{E}\left[\log[X_{m-k,m}]_-^\nu \middle| X_{m-k,m} < C\right] &\overset{d}{=} \mathbb{E}\left[\log[U(Y_{m-k,m})]_-^\nu \middle| Y_{m-k,m} < t_0\right] \\
&\leq \beta^\nu \cdot \mathbb{E}\left[\log\left[Y_{m-k,m} - 1\right]_-^\nu \middle| Y_{m-k,m} < t_0\right].
\end{aligned}$$

Instead of directly showing that the last term is bounded in m, it suffices to show that

$$\mathbb{E}\left[\log[Y_1 - 1]_-^\nu \middle| Y_1 < t_0\right] < \infty.$$

The desired conclusion then follows by stochastic dominance of $Y_{m-k,m}$ over Y_1 near $Y_1 = 1$ for large enough m.

Meanwhile, this last expression is just

$$\int_0^{\min\{1,t_0\}} \frac{(-\log y)^\nu}{(1+y)^2}\, dy \Bigg/ \int_0^{\min\{1,t_0\}} \frac{dy}{(1+y)^2},$$

which can be shown by calculus to be finite for any $\nu > 0$. □

Proof of Lemma 3.1. Using our notation from (5),

$$\begin{aligned}
RBM_{k,n} &= s \cdot (M(s) - M(s-1)) \\
&= s \cdot \binom{n}{s}^{-1} \sum_{\{A \subseteq [n] : |A| = s\}} \log\left(\max_{a\in A}\{X_a\}\right) \\
&\quad - s \cdot \frac{n-s+1}{s} \cdot \binom{n}{s}^{-1} \sum_{\{B \subseteq [n] : |B| = s-1\}} \log\left(\max_{b\in B}\{X_b\}\right) \\
&= \binom{n}{s}^{-1} \sum_{\{A \subseteq [n] : |A| = s\}} \Bigg[s \cdot \log\left(\max_{a\in A}\{X_a\}\right) \\
&\quad - \sum_{\{B \subset A : |B| = s-1\}} \log\left(\max_{b\in B}\{X_b\}\right) \Bigg] \\
&= \binom{n}{s}^{-1} \sum_{\{A \subseteq [n] : |A| = s\}} H_1^{(s)}(X_A),
\end{aligned}$$

since, on the last line, $s-1$ times out of s the largest element in B is the same as the largest in A, and once in s the largest in B is the second largest in A. To obtain the second-to-last line, we used the fact that each set B of size $s-1$ is a subset of $n-s+1$ distinct sets of size s. □

Proof of Lemma 3.2. Without loss of generality, we can assume that the $g^{(n)}$ all have zero mean. By the Efron-Stein ANOVA decomposition [Efron and Stein, 1981], for each $g^{(n)}$, there exist j-parameter symmetric functions $G_j^{(n)}$ with $j = 1, ..., s(n)$ such that

$$g^{(n)}(X_1, ..., X_{s(n)}) = \sum_{j=1}^{s(n)} \sum_{\{I_j \in [s(n)] : |I_j| = j\}} G_j^{(n)}\left(X_{I_j}\right),$$

and the $G_j^{(n)}\left(X_{I_j}\right)$ are all mean-zero and pairwise uncorrelated. Using this result, we can write our U-statistic as

$$U_n = \binom{n}{s(n)}^{-1} \sum_{j=1}^{s(n)} \binom{n-j}{s(n)-j} \sum_{\{I_j \in [n] : |I_j| = j\}} G_j^{(n)}\left(X_{I_j}\right).$$

Moreover, under this notation,

$$G_1^{(n)}(X_1) = \mathbb{E}\left[g^{(n)}(X_1, X_2, ..., X_{s(n)}) | X_1\right],$$

and

$$\widehat{U}_n = \binom{n}{s(n)}^{-1} \binom{n-1}{s(n)-1} \sum_{i=1}^{n} G_1^{(n)}\left(X_i\right). \tag{26}$$

Thus since the G_j are all pairwise uncorrelated and the X_i are *iid*,

$$\begin{aligned}
\mathbb{E}\left[\left(U_n - \widehat{U}_n\right)^2\right] &= \binom{n}{s(n)}^{-2} \sum_{j=2}^{s(n)} \binom{n-j}{s(n)-j}^2 \binom{n}{j} \operatorname{Var}\left[G_j^{(n)}\right] \\
&\leq \frac{s(n)(s(n)-1)}{n(n-1)} \sum_{j=2}^{s(n)} \binom{s}{j} \operatorname{Var}\left[G_j^{(n)}\right] \\
&\leq \frac{s(n)(s(n)-1)}{n(n-1)} \operatorname{Var}\left[g^{(n)}\right],
\end{aligned}$$

which implies the stated result, since $\operatorname{Var}\left[g^{(n)}\right] \leq C$ by hypothesis. □

Proof of Lemma 4.2. We already know from Theorem 3.3 that the finite dimensional distributions of the $X_n(t)$ converge in law to those of $X(t)$. Thus, to show that $X_n(t) \Rightarrow X(t)$ in $\mathcal{D}_{[a,b]}$ for some $0 < a < b$, it suffices by e.g. Theorem 15.6 of Billingsley [1968] to show that there is a constant C such that, given any $\varepsilon > 0$, there exists a constant N_ε such that for all $t_1, t_2 \in [a, b]$ with $|t_1 - t_2| < \varepsilon$ and for all $n \geq N_\varepsilon$,

$$\mathbb{E}\left[\left(X_n(t_2) - X_n(t_1)\right)^2\right] \leq C\varepsilon^2. \tag{27}$$

To show such a bound, it is useful to decompose our expression:

$$\begin{aligned}
\mathbb{E}\left[\left(X_n(t_2) - X_n(t_1)\right)^2\right] \leq\ & \mathbb{E}\left[\left(X_n(t_2) - \widehat{X}_n(t_2)\right)^2\right] \\
&+ \operatorname{Var}\left[\widehat{X}_n(t_2) - \widehat{X}_n(t_1)\right] \\
&+ \mathbb{E}\left[X_n(t_2) - X_n(t_1)\right]^2 \\
&+ \mathbb{E}\left[\left(\widehat{X}_n(t_1) - X_n(t_1)\right)^2\right],
\end{aligned}$$

where $\widehat{X}$ is a Hájek projection of X as defined in (10). It now remains to bound the terms individually.

By Lemma 3.2, the first and the last summands decay uniformly as $O(1/k(n))$ on $[a, b]$, and so become eventually negligible for any $\varepsilon > 0$. Meanwhile, as in (14), we can write the variance term (i.e. the second summand) as

$$\frac{4n}{t_1 t_2 k(n)} \operatorname{Var}\left[\mathbb{E}\left(H_1^{\left(\frac{2n}{k(n)t_2}\right)} - H_1^{\left(\frac{2n}{k(n)t_1}\right)} | X_1\right)\right],$$

which by Lemma 8.3 converges to

$$\frac{\gamma^2 (t_2 - t_1)^2}{t_1 t_2 (t_1 + t_2)} \leq \frac{\gamma^2}{2a^3}(t_2 - t_1)^2$$

on $[a, b]$; the result can be extended to show that the convergence is uniform over the interval. Finally, Lemma 8.2 reduces the problem of showing that $\mathbb{E}[X_n(t)]$ satisfies the required

property to showing that $a_n(t) = A(2n/tk(n))$ satisfies it; this latter task can be performed using the Potter bounds. Thus (27) holds. □

References

J. Beirlant, G. Dierckx, A. Guillou, and C. Stărică. On exponential representations of log-spacings of extreme order statistics. *Extremes*, 5(2):157–180, 2002.

J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of extremes: theory and applications*. John Wiley & Sons Inc, 2004.

P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, 1968.

S. Coles. *An introduction to statistical modeling of extreme values*. Springer Verlag, 2001.

S. Csörgő, P. Deheuvels, and D. Mason. Kernel estimates of the tail index of a distribution. *The Annals of Statistics*, 13(3):1050–1077, 1985.

J. Danielsson, L. de Haan, L. Peng, and C.G. de Vries. Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate analysis*, 76(2):226–248, 2001.

L. de Haan and A. Ferreira. *Extreme value theory: an introduction*. Springer Verlag, 2006.

A. Dekkers, J. Einmahl, and L. de Haan. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, pages 1833–1855, 1989.

H. Drees. On smooth statistical tail functionals. *Scandinavian Journal of Statistics*, 25(1): 187–210, 1998.

H. Drees and E. Kaufmann. Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications*, 75(2):149–172, 1998.

H. Drees, L. de Haan, and S. Resnick. How to make a Hill plot. *The Annals of Statistics*, 28(1):254–274, 2000.

B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596, 1981.

A. Feuerverger and P. Hall. Estimating a tail exponent by modelling departure from a Pareto distribution. *The Annals of Statistics*, 27(2):760–781, 1999.

M. Gomes, L. de Haan, and L. Rodrigues. Tail index estimation for heavy-tailed models: accommodation of bias in weighted log-excesses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):31–52, 2008.

A. Guillou and P. Hall. A diagnostic for selecting the threshold in extreme value analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):293–305, 2001.

P. Hall. Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of Multivariate analysis*, 32(2):177–203, 1990.

P. Hall and AH Welsh. Adaptive estimates of parameters of regular variation. *The Annals of Statistics*, 13(1):331–341, 1985.

B. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, pages 1163–1174, 1975.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

P. Loève. Fonctions aléatoires du second ordre. In P. Lévy, editor, *Processus stochastiques et mouvement Brownien*. Gauthier-Villars, 1948.

D. Mason and T. Turova. Weak convergence of the Hill estimator process. *Extreme Value Theory and Applications*, pages 419–432, 1994.

J. Petty, S. de Treville, and S. Wager. Economies of extremes: The case of venture capital decision making. Submitted, 2012.

H. Potter. The mean values of certain Dirichlet series, II. *Proceedings of the London Mathematical Society*, 2(1):1–19, 1942.

A. Rényi. On the theory of order statistics. *Acta Mathematica Hungarica*, 4(3):191–231, 1953.

S. Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling.* Springer Verlag, 2007.

S. Resnick and C. Stărică. Smoothing the Hill estimator. *Advances in Applied Probability*, pages 271–293, 1997.

Department of Statistics, Stanford University, CA-94305 Stanford, USA.
E-mail address: swager@stanford.edu